BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25852
doi: 10.3897/biss.2.25852

OPEN ACCESS

Conference Abstract

Aggregation and Synthesis of Taxon Information in the Encyclopedia of Life

Katja S Schulz‡, Jennifer Hammock‡,‡

‡ Smithsonian Institution, Washington, United States of America

Corresponding author: Katja S Schulz (schulzk@si.edu)

Received: 15 Apr 2018 | Published: 03 Jul 2018

Citation: Schulz K, Hammock J (2018) Aggregation and Synthesis of Taxon Information in the Encyclopedia of Life. Biodiversity Information Science and Standards 2: e25852. https://doi.org/10.3897/biss.2.25852

Abstract

To improve access to biodiversity knowledge for diverse audiences, the Encyclopedia of Life (EOL) aggregates materials from hundreds of content providers. In addition to text, media, references, taxon names and hierarchies, traits and other structured data are an increasingly important component of EOL (TraitBank). Content priorities for TraitBank include information about body size, geographic distribution, habitat, trophic ecology, and biotic interactions in general. Our goal is to summarize available data at the level of species and supraspecific taxa and to achieve broad taxonomic coverage for high priority topics. Integration of information from heterogeneous sources relies on a variety of community standards (e.g., Dublin Core, Darwin Core, Audubon Core) as well as post-hoc semantic annotations that standardize terminology for traits and metadata and provide links to domain ontologies and controlled vocabularies (e.g., Ontology of Biological Attributes, Phenotypic Quality Ontology, Environment Ontology, Uber Anatomy Ontology). Taxon names are mapped to a reference hierarchy that leverages taxonomic information from many different resources (e.g., Catalogue of Life, World Register of Marine Species, Paleobiology Database, National Center for Biotechnology Information). Names reconciliation takes into account canonical name strings, authorities, and synonym relationships as well as information about ranks and hierarchies (parent/child taxa). In EOL version 3 this infrastructure supports complex queries across EOL data sets, autogenerated natural language descriptions of taxa, and knowledge-based recommender systems for the exploration of content along multiple axes, including phylogeny, ecology, life

© Smithsonian Institution. This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

history, relevance to humans and other characteristics derived from structured data. Most TraitBank data currently come from published data compilations and databases of specialist projects, but there are still significant gaps in coverage for many lesser known groups. Recent advances in natural language processing, image analysis, and machine learning technologies, facilitate the automated extraction and processing of data from unstructured text and images. This will soon make it possible to recruit vast amounts of information from millions of pages of taxonomic, ecological, and natural history literature available in open access repositories like Biodiversity Heritage Library (BHL) and Plazi. Natural history collections are another promising source of new taxon information. Millions of museum specimens indexed by organizations like the Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections (iDigBio) already contribute significantly to our understanding of species occurrences in space and time. But specimens and associated labels and field notes can also provide information about morphology, phenology, habitats, and biotic interactions. Data mined from literature corpora or specimen collections will generally lack detailed descriptions of what exactly was measured, metadata about the data capture process, measurement accuracy, and other important parameters. The integration of this information with data sets from the primary literature therefore poses challenges that go beyond the standardization of taxonomy and terminology. Leverage of data from a wide variety of sources is however necessary to achieve a comprehensive, interconnected biodiversity knowledge base that supports the exploration of trait diversity across the tree of life.

Presenting author

Katja S Schulz